DOCUMENT RESUME

ED 173 797 CS 204 941

AUTHOR Brodkey, Linda; Young, Rodney W.
TITLE A Sensible, Interesting, Organized, Rhetorical Procedure for the Grading of Student Essays.
PUB DATE Apr 79
NOTE 18p.; Paper presented at the Annual Meeting of the Conference on College Composition and Communication (30th, Minneapolis, Minnesota, April 5-7, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Composition (Literary); English Instruction; *Evaluation Criteria; *Grading; Higher Education; Persuasive Discourse; Statistical Analysis; *Student Evaluation; Teacher Attitudes

ABSTRACT

A study was conducted to test the reliability of a college English teacher's criteria for grading papers of argumentative prose. Information was sought regarding the relation of marginal and end-of-paper comments to the grade itself, as well as of the ability of multiple regression analysis to express the relationship of the individual and collective contributions of the variables to the grades. A group of 36 previously graded student essays were analyzed for grading variables; 14 such variables emerged. The papers were reevaluated on relative performance for each variable. The four variables found to account significantly for the grade were those concerned with interest in topic, scope, organization, and rhetoric. (DF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Linda Brodkey
Rodney W. Young
Testing Division
The University of New Mexico
Albuquerque, New Mexico 87131

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Linda Brodkey
Rodney W. Young
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED173797

A Sensible, Interesting, Organized, Rhetorical Procedure for the Grading of Student Essays

Teachers are actors. None more so than composition teachers. We are actors in the sense that the teaching of writing requires as much artifice as art, a persuading of our students and ourselves that we know what to do. Within limits of course we do know a great deal about writing. We know that some relation must obtain between what we can identify and describe in writing--rhetoric, grammar, spelling, punctuation--and writing itself. What we do not know is the precise character of the relationship or the ways in which this relationship can be explained so that students will invariably write better prose. In the hope that the value of such discrete skills will be revealed to students, more or less in practice, we ask them in the course of a semester to write a number of essays, in which it is understood that at least some of these writing skills will be displayed. And, it is in grading these essays that most of us confront not only what our students do not understand about writing, but also what we ourselves do not know.

CS204 941

The very act of grading, actually assigning a grade to a piece of prose, generally means changing roles. As teachers we may be supportive, cheerful, optimistic, and generous, but when we grade we are judges. A good set of essays reflects the success of both students and teacher. We are inclined to see a poor set, however, as a direct result of poor teaching. Teachers are never so vulnerable as when they grade their students' essays. If the teaching of writing requires a bit of deception, so that students will work at writing, the grading of their efforts often amounts to an unexpected, even an unwanted, flood of house lights, which not only exposes, but accentuates our fragile knowledge of pedagogy. All insecurities, intellectual and emotional, seem revealed in grading. Are the grades fair; do the same criteria apply to all students; are the criteria we use the ones we think we use; are some students' papers read more rigorously than others? At times, the list of self-recriminations is endless and the sensation that we really do not know what we are doing overwhelming.

We can all expect to learn a great deal more about grading in the next few years. Already we have Paul B. Diederich's study of the grading practices of the Educational Testing Services graders in Measuring Growth in English (Urbana: NCTE, 1974), pp. 6-10, in which he mentions five kinds of criteria used by those graders. And, we have E. D. Hirsch, Jr.'s argument in The Philosophy of Composition (Chicago: University of Chicago Press, 1977), pp. 178-188, which begins with Diederich's inductive conclusions and proceeds along the

lines of deduction, concluding that grading can be classified as two philosophies, intrinsic and extrinsic judgments--each of which represents different teaching and writing goals--and only one of which, intrinsic judgment, Hirsch himself cares to defend. There is much to be learned from what graders might agree to value in the assessment of student writing; all of us need to know what we might reasonably expect of ourselves and each other as graders of student prose. There remains, however, a strong possibility that learning about what others do as graders will only increase the practicing teacher's insecurities about his or her actual grading practice. It is one thing to know what research studies recommend, and entirely another matter to know whether or not you are yourself a reliable grader, whose criteria could be explained as reasonable, consistent, and a valid explanation of the grade. Most of us will not be participating in grading studies because such research is mostly very expensive and will be conducted largely at colleges and universities where money and time will be appropriated for the express purpose of determining criteria. This does not mean, however, that we are destined to know nothing more about what we do as graders than we can learn by reflecting on our activities or by reading the results of large studies when they are made available to us. There are relatively inexpensive ways to discover a great deal about your own criteria and the ways in which you apply them in grading, if your school has an up-to-date computer and you are willing to subject yourself to a rather dispassionate statistical analysis of what you actually do.

The grading project reported in this essay is such an inexpensive project, and it was motivated almost entirely by the composition teacher's suspicion that her own grades would not stand up in court, that is, that she would not care to defend them in public. Nonetheless, she felt that she had a need and her students a right to know more about her reliability as a grader, and she also wished to know as much as possible about the criteria themselves. To the extent that the criteria created for and used in this study recommend themselves not only as reflections of a teacher's experience and intuition, but also as the internally reliable and valid variables which account for the grades she assigns essays, we would argue that the variables themselves merit more attention than we give them here. Our main purpose in reporting our results is not, however, a defense of these particular variables, but rather a presentation of the technique used in verifying and validating their part in her grading, a technique we consider to be equally useful to others, whose interest in their own grading criteria and practices approximate those of the teacher in this study. We were guided by two assumptions. We assumed that we could expect a relation between what she wrote in explanation of the grade (end-of-paper and marginal comments) and the grade itself. Moveover, we assumed that multiple regression analysis, because of its ability to relate and order the contribution of independent variables to a dependent variable, would express the relationship in terms useful to teachers, namely, as a statistical account of both the individual and collective contribution of the variables to the grade.

The 36 argumentative essays used in the study were collected by the classroom teacher in the tenth week of a one semester English 101 course at the University of New Mexico. The course aims at teaching expository writing skills to all entering students whose ACT English scores are 19 or better and to students with less than 19 on ACT English, who have successfully completed English 100, a remedial writing course intended to prepare them for English 101. The teacher read and graded the essays, making marginal and end-of-paper comments in her customary way. Having decided that she would reserve this set of essays for later study, she copied the graded set before returning them to the students. Of the 41 essays turned in and graded, 36 were judged suitable for the study, which began in earnest some twenty weeks later. The five essays not used in the project were flawed in copying.

First, we combed the marginal and end-of-paper comments for the purpose of constructing the teacher's grading variables. For instance, the comment, "I think your argument would have been more effective if you had included an example in defense of your position," became the variable, Evidence, "supports generalization with pieces of evidence." Such comments and subsequently such variables, we felt, represented categorical grading criteria, that is, constructs that we believed might very well explain the grade to the students and should, therefore, serve as variables in the multiple regression analysis. We made no special attempt, and have made none here, to transform the

teacher's comments into prose more like that used in student handbooks, preferring instead to retain as much as possible the sense and language actually used by the teacher in commenting to a group of students she had worked with for ten weeks. Nor have we disguised the fact that some variables, in particular, Quality of Evidence and Thesis/Body Match, bore such a superficial resemblance to each other that even we doubted that each made a unique contribution to the grade. However, we reasoned that the grader's tendency to express what appeared to be similar kinds of information in different terms might signify a substantive difference in concept. In all, we constructed fourteen variables, all of which are listed and described below:

Clear -- reader can understand the points the writer is making

Interesting -- the topic strikes the reader as significant

Sensible -- the scope of the essay is reasonable, neither too large nor too small

Point of View -- a sense that a thoughtful being wrote the essay

Rhetoric -- uses devices of expansion, e.g., redundancy or metaphor

Organization -- some apparent principle of arrangement, i.e., time, space, relative order of importance

Grammar -- generally acceptable written English, conventional spelling and punctuation

Usage -- careful attention to diction, fluency

Evidence -- supports generalization with pieces of evidence

Quality of Evidence -- the evidence is suited to the generalization

Logical Transition -- the ability to move from one sentence to the next and from one paragraph to the next

Thesis/Body Match -- the evidence and generalization seem suited to one another, i.e., evidence supports that generalization and not another

Not an Obvious Argument -- the argument seems worth the reader's time, i.e., not about something everyone already understands or already proven

Fair -- considers the possibility of opposing points of view

Having agreed that the variables expressed the sense of her comments, the teacher then reevaluated the essays, this time rating each on its relative performance in each category (variable). We used a Likert scale: one indicating a low level of performance and five, a high level. The Likert scores, along with the grade assigned each essay during the term, now converted from a letter grade to a numerical score ranging from 12 to 1 (to incorporate pluses and minuses), were entered into a multiple regression equation. We used traditional step-down procedure wherein the computer simultaneously enters all variables into the equation on the first step. The overall results of this analysis were statistically significant, warranting testing of the contribution of the individual variables ($F=10.990$, $p<.01$). The multiple R was .931 and R^2 was .867.

We are fully aware of the injunction against using multiple regression when the group is smaller than 50 and the ratio of variables to subjects slightly less than three to one (or in some instances, ten to one). We understand also that the R^2, .867, is inflated because of the variables-to-subjects ratio, although significant. We reasoned, however, that since multiple regression is essentially analysis of variance, a procedure routinely applied to groups as small as ours, our use of the method could be justified if the final results were conclusive.

Once we had determined that the overall regression was significant, we then analyzed the significance of the individual variables. The results of this second analysis are depicted in Table I, from which we learned that Interesting, Sensible, Organization, and Rhetoric significantly account for the grade:

Table I

Fourteen Variables Initially Examined for Grade Prediction

Variable	β	t	df	P
Not an Obvious Argument	.1304	n/s	—	—
Clear	-.0364	n/s	—	—
Evident	-.1264	n/s	—	—
Fair Argument	-.0360	n/s	—	—
Grammar	-.1597	n/s	—	—
Interesting	.2894	1.906	22	<.07*
Thesis/Body Match	-.0144	n/s	—	—
Organization	.3778	2.131	22	<.04
Quality of Evidence	-.0470	n/s	—	—
Rhetoric	.5043	2.789	22	<.01
Sensible	.3239	2.094	22	<.05
Point of View	——	—— **	—	—
Logical Transition	.0960	n/s	—	—
Usage	-.1724	n/s	—	—

* INT left in analysis even though slightly higher than a p<.05 because of its subsequent high degree of statistical significance

** Insufficient contribution to use; significance tests based on 13 variables.

Since Fred N. Kerlinger and Elazar J. Pedhazur in Multiple Regression in Behavioral Research (New York: Holt, Rinehart and Winston, 1973), pp. 446-47, do not recommend multiple regression for small groups because of the bias in the regression statistics, we needed to find out if these four variables, when used alone, would strongly account for the grade. This indeed was the case. We feel that the verification of Interesting, Organization, Sensible, and Rhetoric by the equation justifies our procedure, as well as suggests the suitability of the procedure for other research projects in which the selection of variables is tempered by the experience of the researchers. In other words, given the tradition of end-of-paper comments, we had good reason to expect that among the variables we had constructed at least some were important predictors of the grade. If the equation had identified no such variables, of course we would have abandoned the project.

The overwhelming significance of Interesting, Organization, Sensible, and Rhetoric on the second multiple regression analysis, when only these four variables and the grade were used, was sufficiently strong to consider the results valid. Using all 14 variables (actually only 13; see second note in Table I), we accounted for 87% of the grade; in the second analysis, using only four variables, we accounted for 83% of the grade. The 87% is inflated by the variables-to-subjects ratio to a far greater extent than the 83%, demonstrating that the 83% is actually a larger number in real terms than the 87%. Once the insignificant variables are out of the

equation, each of the four variables rises in significance, and there is a decreasing probability that the relation between the grade and the variables is coincidental:

Table II

The Four Significant Variables Related to Grade Prediction

Variable	β	t	df	P
Interesting	.3018	3.041	31	<.01
Organization	.2511	2.504	31	<.02
Rhetoric	.3158	3.176	31	<.01
Sensible	.3173	3.092	31	<.01

We should also like to point out that the manner in which the four variables account for the grade is not cumulative. In other words, Interesting does not predict a certain percentage, Rhetoric a portion, Organization a bit more, and Sensible whatever remains. Instead, the 83% represents the combined weight of these variables, what they account for as a cluster of interlocking constructs. These variables are inextricably linked one to the other, and a student's relative performance in all four predicts the grade.

The predictive power of the cluster is also significant both to teachers and researchers because we are satisfied that even though they combine to form such a powerful basis of evaluation, each variable is distinct from the others; Interesting is not a synonym for Rhetoric, or Organization, or Sensible. Further, as shown below in Table III, the results of our commonality analysis, no single variable, no two variables, no combination of any three variables approaches the predictive strength of the quartet:

Table III

Commonality Analysis of Four Contributing Variables and Their Combinations

Variable(s)	Percentage
Interesting	.05034
Sensible	.03413
Rhetoric	.05490
Organization	.05205
Int + Sns	.12283
Int + Rht	.00164
Int + Org	.01449
Sns + Rht	.00716
Sns + Org	.01120
Rht + Org	.16759
Int + Sns + Rht	.01556
Int + Sns + Org	.05425
Int + Rht + Org	.03050
Sns + Rht + Org	.04172
Int + Sns + Rht + Org	.17290
Total	.83126

Their aggregate grade prediction, taken in tandem with their expression of unique grading constructs, provides a statistical sketch of what many practicing teachers would argue on intuitive grounds; that writing is a display of thorough-going or holistic language use, not a matter of discrete skills, such as putting commas in the right places, which is all too often what text books, students, and we ourselves are inclined to make of it. What multiple regression portrays in this instance is a lively juggling of four intellectual constructs, whose heuristic is clearly rhetorical, for terms like Interesting, Sensible, Organization, and Rhetoric are informed by the tradition of rhetoric not grammar. It is altogether likely, however, that the four significant variables, balanced both by writer and grader, have absorbed at least some of the information represented by the variables excluded in the first multiple regression analysis. We would speculate that the teacher in judging an argument to be Interesting, Sensible, Organized, and Rhetorical, according to her scale of relative performance, is concommitantly evaluating its use of punctuation, its spelling, its grammar. In short, we imagine she encorporates some of these discrete skills into her ideas of what it means to be an Interesting, Sensible, Organized, and Rhetorical writer.

Given the procedures we followed in testing the validity of the grade and end-of-paper comments, we were easily able to establish the teacher's reliability as a grader of the argumentative essays used in the study. The teacher assigned the grade in the tenth week of the semester. We constructed the grading variables and tested

their validity in the course of our study, which began some twenty weeks after the essays themselves were graded. In effect, then, we recreated the conditions for a test-retest reliability study, a natural consequence, in this instance, of the validity study, for we needed the grade to work as the dependent variable in the multiple regression equation. Each essay, therefore, has been graded twice -- once during the semester and once according to the variables constructed for the study.

Thus, the multiple R of .912, by whose square we earlier expressed the degree (percentage) to which the four significant variables predict the grade, can also be read as the reliability coefficient for the grading. Reliability, based on test-retest, is the correlation coefficient; thus the grader is reliable at .912, an impressive reliability coefficient in statistical or practical terms. The teacher can be relatively certain that she consistently bases most of the grade on the four significant variables.

By way of summary, we would like to comment briefly on some implications of studying the grading practices of classroom teachers. First, it is some comfort for the teacher, whose grading criteria were used in this study, to know that she is consistently paying close attention to four variables, in her case, Interesting, Sensible, Organization, and Rhetoric, for like many writing teachers, she suspected that her criteria were erratic or whimsical. Knowing the overwhelming importance she attaches to these variables, she now makes special attempts to relate her pedagogy to the criteria.

That we do not yet know much about the precise configuration of the constructs being represented by the variables clearly suggests that in future studies we attempt to specify her perception of Interesting, Sensible, Organized, Rhetorical essays -- perhaps by path analysis And, we also hope that our results will encourage other teachers to describe their own grading practices by replicating our procedure or by devising techniques of their own. Before any of us can confidently recommend what the criteria for evaluating writing should be, it is first necessary to describe present criteria, the pedagogical considerations they express, and how consistently we might expect a classroom teacher to apply them in assessing student work.